# Perceptual Scale-Space and Its application

Yizhou Wang[1], Siavosh Bahrami[1] and Song-Chun Zhu[1,2]

Departments of Computer Science[1] and Statistics[2]
University of California, Los Angeles
Los Angeles, CA 90095
email: {wangyz, siavosh, sczhu}@cs.ucla.edu

**Abstract**

When an image is viewed at decreasing resolutions in a Gaussian pyramid, information is lost gradually. Amid continuous intensity changes across scales, there are "quantum jumps" or "perceptual transitions" in our inner representation. In this paper, we study a representational paradigm called the perceptual scale-space which augments the Gaussian pyramid in traditional image scale-space theory by constructing a so-called sketch pyramid. Each level of the sketch pyramid is a generic attribute graph – called primal sketch, and is inferred from the corresponding image at the same level of a Gaussian pyramid by Bayesian inference using a generative image model. Perceptual jumps are then represented by structural changes in the primal sketch in terms of graph operators, such as death-birth and split-merge of vertices and edges in the generic attribute graph. In a training stage, we ask seven human subjects to label transitions of graphs over scales for a set of images. We learn the most frequent atomic graph operators, composite operators, and thresholds of transitions across human subjects. This information is then used in a generative model for inferring a sketch pyramid up-down a Gaussian pyramid. In experiments, we show that the sketch pyramid is a more parsimonious representation than a multi-resolution wavelet transforms, and demonstrate an application on adaptive image display – showing a large image in a small screen (say PDA) through a selective tour of its image pyramid. In this application, the sketch pyramid provides a means for calculating information gain in zooming-in different areas of a image by counting a number of operators expanding primal sketches, such that the maximum information is displayed in a given number of frames. We argue that the perceptual scale-space enriches the conventional scale-space theory, and provides an important representation for many vision applications, such as super-resolution and multi-scale object recognition.

**Keywords**

Scale-Space, Image Pyramid, Primal Sketch, Stochastic Graph Grammar, Generative Modeling.

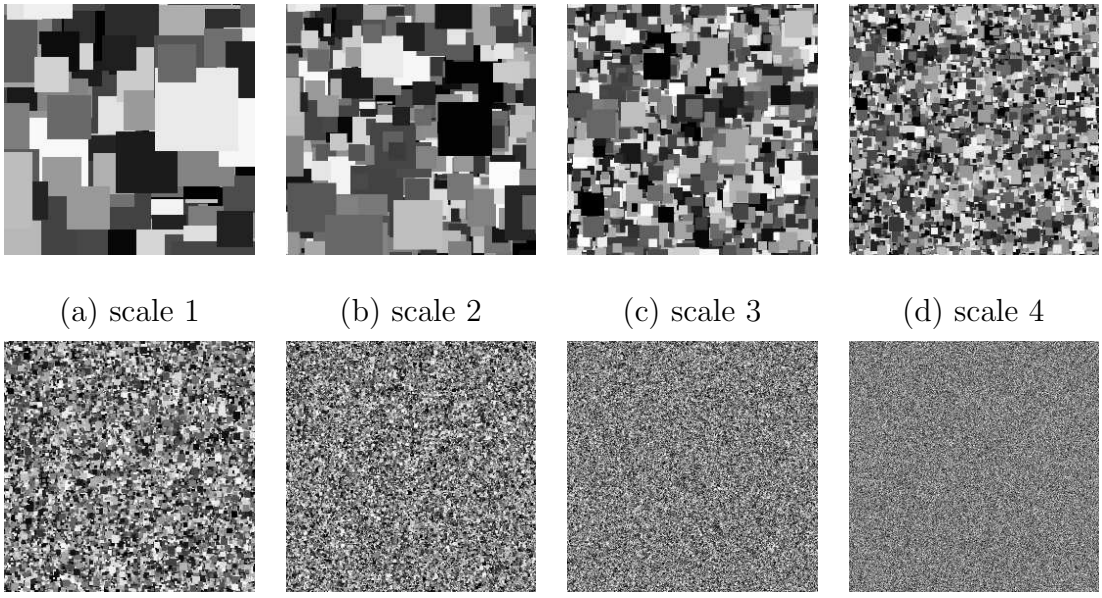
An early version of this paper appeared in ICCV05 as an oral presentation. Submitted to IJCV.

## I. Introduction

It has long been noticed that objects viewed at different distances or scales may create distinct visual appearances. As an example, Figure 1 shows trees in a long range of distances. In region A at near distance, the shapes of leaves can be perceived. In region B, the image becomes more complex, and we cannot see individual leaves. Instead we perceive a collective foliage impression. In region C at an even further distance, the image becomes stochastic texture. Finally, in region D at a very far distance, the image appears to be a smooth region whose intensities, if normalized to $[0, 255]$, are independent Gaussian noise.


Fig. 1. A scene with trees at a large range of distances. Leaves at regions A, B, C, D appear as shape, texture, and noise, respectively.

These perceptual changes become more evident in a series of simulated images shown in Figure 2. We simulate the leaves by uniform intensity squares over a finite range of sizes. Then we zoom out the images by a $2 \times 2$-pixel averaging and down-sampling process, in a way similar to constructing a Gaussian pyramid. The 8 images are the snapshots of the image at 8 consecutive scales. At high resolutions, we see edges, boundaries, and corners. At middle resolutions, these geometric elements disappear gradually and appear as texture. At the lowest resolution, each pixel at scale 8 is the averaged sum of $128 \times 128$ pixels in scale 1 and covers many independent "leaves." Therefore, the image becomes Gaussian noise according to the central limit theorem in statistics.



(a) scale 1 (b) scale 2 (c) scale 3 (d) scale 4

(e) scale 5 (f) scale 6 (g) scale 7 (h) scale 8

Fig. 2. Snapshot image patches of simulated leaves taken at 8 consecutive scales. The image at scale $i$ is a $2 \times 2$-pixel averaged and down-sampled version of the image at scale $i-1$. The image at scale 1 consists of a huge number of opaque overlapping squares whose lengths fall in a finite range $[8, 64]$ pixels and whose intensities are uniform in the range of $[0, 255]$. To produce an image of size $128 \times 128$ pixels at scale 8, we need an image of $128^8 \times 128^8$ at scale 1.

In this paper, our objective is to study the "quantum jumps" or "perceptual transitions" in image representations or vision models amid the continuous intensity changes across scales. More specifically, we are focused on generic perceptual transitions at low-middle level vision rather than object specific perceptions at high level vision.

We approach the problem from an information theoretical viewpoint and seek for an efficient representation over scales in terms of minimum coding length or maximum a posterior probability. We shall use simple experiments with human subjects to set parameters or thresholds for the perceptual transitions, but it is beyond the scope of this paper to link these transitions directly to their potential biologic base on the psychophysical or neurophysiological levels.

To account for perceptual transitions, we propose a new representational paradigm called

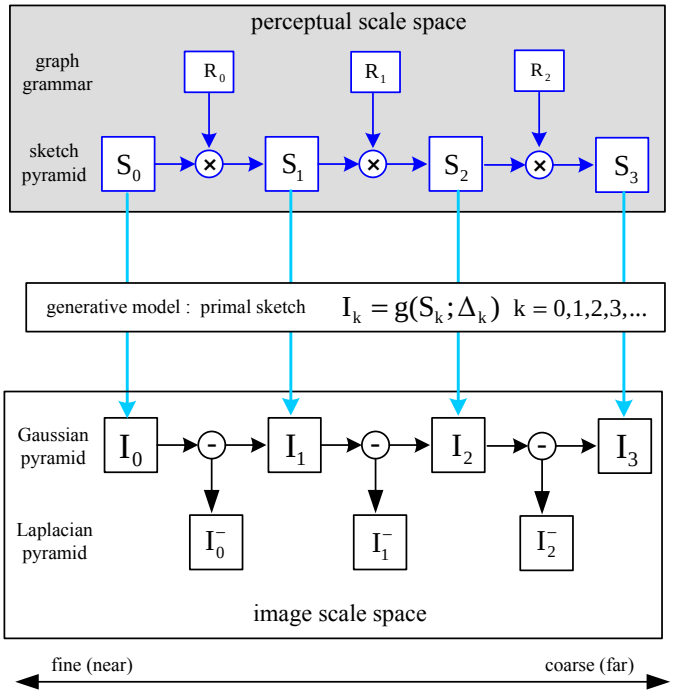



Fig. 3. The diagram of a perceptual scale-space representation. It consists of a pyramid of primal sketches represented as attribute graphs and a series of graph operators. These operators are represented as graph grammar rules for perceptual transitions. The primal sketch at each level generates an image in a Gaussian pyramid using a dictionary of image primitives.

*perceptual scale-space* shown in Figure 3. It has the following two components.

The first is a pyramid of primal sketches. Primal sketch is a concept conjectured by Marr [18] as a generic "token representation" for early vision. According to Marr, it should be parsimonious and sufficient to reconstruct the image, and the basic tokens are image primitives, such as edge elements, bars, corners, blobs, and terminators, similar to the textons proposed by Julesz [9]. A rigorous mathematical model of the primal sketch was proposed in [6], [8] which includes two layers of Markov random fields. The top layer is an attribute graph for image structures using a dictionary of image primitives and is called the Gestalt field, and the bottom layer is a Markov random field for stochastic textures with the Gestalt field being its boundary condition. As it is shown by the downward arrows in Figure 3, each primal sketch in the sketch pyramid generates an image in the Gaussian pyramid by a

generative model. We shall discuss this generative model in a later section.

The second component of perceptual scale-space is a series of graph operators. These operators are represented as graph grammar rules for the structure (topology) changes of the attribute graphs between two levels in the sketch pyramid, such as split-merge, death-birth of vertices or edges. These are qualitative perceptual jumps to account for generic perceptual transitions over scales.

In comparison, the image representations adopted in classical scale-space theory [31], [12], [13], [23] are Gaussian and Laplacian pyramids. A Laplacian pyramid consists of band-passed images which are the difference between every two consecutive levels of images in a Gaussian pyramid. These band-passed images show intensity changes over scales and can be further decomposed into wavelet elements (e.g. Gabors). Gaussian and Laplacian pyramids have dominated image representation and coding in low-level vision and image processing. However, they are restricted to the description of raw pixel intensities and their changes in a deterministic setting. Even though the classical scale-space theory studied discrete and qualitative events, such as appearance of extremal points [31], and tracking inflection points, such descriptions are deterministic functions of images. Thus we call them image scale-space in contrast to perceptual scale-space.

One may view perceptual scale-space as an augmented representation to image scale-space. A sketch pyramid is inferred from a Gaussian pyramid by maximum posterior probabilities to yield a generic perceptual representation. A series of grammar rules account for qualitative changes of sketches derived from Laplacian images. In perceptual scale-space, each object, part, or primitive has a finite "lifespan" – a range of scales that they can infer.

In this paper, the following three issues are studied for perceptual scale-space.

1. We identify three categories of perceptual transitions in perceptual scale-space. (i) Blurring of image primitives without structural changes. (ii) Graph grammar rules for graph topological changes. (iii) Catastrophic changes from structures to texture with massive im-

age primitives disappearing at certain scales. For example, the individual leaves disappear from region A to B.

2. In our experiments, we find the top twenty most frequent occurring graph operators and their compositions. As the exact scale at which a perceptual jump occurs may vary slightly from person to person, we asked seven human subjects to identify and label the perceptual transitions over the Gaussian pyramids of fifty images. The statistics of this study is used to decide parameters in the Bayesian formulation for the perceptual transitions.

3. We infer the sketch pyramid in the Bayesian framework using learned perceptual graph grammar rules, and compute the optimal representation upwards-downwards the pyramid, so that the attribute graphs across scales are optimally matched and have consistent correspondence. We call the result the "perceptual sketch pyramid". Experiments are designed to verify the inferred perceptual scale-space by comparing with Gaussian/Laplacian scale-space and human perception.

4. We demonstrate an application of perceptual scale-space: adaptive image display. The task is to display a large high-resolution digital image in a small screen, such as a PDA, a windows icon, or a digital camera viewing screen. Graph transitions in perceptual scale-space provide a measure in term of description length of perceptual information gained from coarse to fine across a Gaussian pyramid. Based on the perceptual information, the algorithm can decide to show different areas at different resolutions so as to convey maximum information in limited space and time.

Other potential applications of perceptual scale-space include super-resolution [33] and multi-resolution object recognition [34].

The perceptual scale-space representation is related to a range of work in computer vision literature besides its close ties to the classical scale-space theory [31], [12], [13], [23].

First, it has long been acknowledged that certain image features/structures only exist within a small range of scales [31]. Therefore, pursuit of scale invariant features [14], [10],

[20] in object recognition is valid only in a small range of scales. Existing work addresses this problem mostly from a discriminative perspective by designing stable feature detectors. We take a generative method and study perceptual transitions and their compositions of image features explicitly.

Second, there is also abundant work in the literature studying natural image statistics (see a survey paper in [26]). One fundamental observation in natural images is the scale-invariant properties [21], [22], [36]. These invariant statistics are extracted from entire images. For example, power spectrums of the Fourier transform or histograms of gradients. Such global statistics are often invariant over a range of scales, with a common interpretation that natural images contain objects of various sizes [21]. The perceptual scale-space theory is not contradictory to the scale invariance observations. Our study is focused on perceptual changes of specific objects rather than the global image statistics. The drastic transitions of images, shown in Figure 2, is because such image patterns differ from natural images in two respects. (i) The squares have a smaller range of sizes than objects in natural images. (ii) The square images have 8 scales, but natural images usually can be only scaled and down-sampled 4 to 5 times.

Third, the work is closely related to a number of early papers in the authors' group. This work is a direct extension of the primal sketch work[6], [7] to multi-scale. It is also related to the study of topological changes in texture motion [29]. In a companion paper[32], we studied the theoretical issues of entropy rate changes over scales and compare the two regimes on image models: wavelet and sparse coding for structures and Markov random fields for stochastic textures. This paper is mostly focused on generic mild topological transitions over scales, in contrast to the class specific transition in [34], where a detailed multi-scale modeling of specific human faces is studied.

The paper is organized as follows. First, we set the background by introducing the primal sketch work and compare it to image pyramids and sparse coding in Section II. Second, we

introduce the perceptual uncertainty and three types of transitions in Section III. Then the perceptual scale-space theory is studied in three sections – Section IV presents a generative model representation, Section V discusses the learning issues of a dictionary of graph operators and their parameters, and Section VI introduces an inference algorithm in the Bayesian framework with some experiment results of computing the sketch pyramids. Then we show the application of the sketch pyramids in adaptive image display in comparison with image pyramid methods. The paper is then concluded with a discussion in Section VIII.

## II. Background: image pyramids and primal sketch

This section briefly reviews primal sketch and image pyramid representations and compares their performance in image reconstruction to provide background information.

### *A. Image pyramids and sparse coding*

Gaussian and Laplacian pyramids are the central representation in classical scale-space theory. A Gaussian pyramid is a sequence of images $\{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_n\}$. Each image $\mathbf{I}_k$ is computed from $\mathbf{I}_{k-1}$ by Gaussian smoothing (low-pass filtering) and down-sampling.

$$\mathbf{I}_k = \lfloor G_\sigma * \mathbf{I}_{k-1} \rfloor, \ k = 1, 2, ..., n, \quad \mathbf{I}_0 = \mathbf{I}.$$

A Laplacian pyramid is a sequence of images $\{\mathbf{I}_1^-, \}$. Each image $\mathbf{I}_k^-$ is the difference between an image $\mathbf{I}_k$ and its smoothed version,

$$\mathbf{I}_k^- = \mathbf{I}_k - G_\sigma * \mathbf{I}_k, \ k = 0, 2, ..., n-1.$$

This is equivalent to convolving image $\mathbf{I}_k$ with a Laplacian (band-pass) filter $\Delta G_\sigma$. The Laplacian pyramid decomposes the original image into a sum of different frequency bands. Thus, one can reconstruct the image by a linear sum of images from the Laplacian pyramid,

$$\mathbf{I} = \mathbf{I}_0, \mathbf{I}_k = \mathbf{I}_k^- + G_\sigma * \lceil \mathbf{I}_{k+1} \rceil, k = 0, 1, ..., n-1.$$

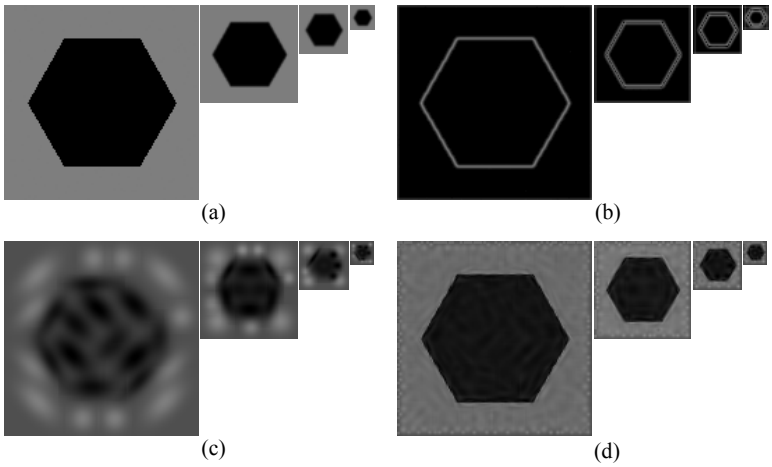



Fig. 4. The image pyramids for an hexagon image. (a) The Gaussian pyramid. (b) The Laplacian of Gaussian pyramid. (c) Reconstruction of (a) by 24, 18, 12, 12 Gabor bases respectively. (d) Reconstruction of (a) by Gabor bases. The number of bases used for a reasonable satisfactory reconstruction quality is 500, 193, 80, 38 from small scale to large scale respectively.

The down-sampling and up-sampling rates will be decided by $\sigma$ in accordance with the well-known Nyquist theorem (see [16] for details).

One may further decompose the original image (or its band-pass images) into a linear sum of independent image bases in a sparse coding representation. We denote $\Psi$ as a dictionary of over-complete image bases, such as Gabor cosine, Gabor sine, and Laplacian bases.

$$\mathbf{I} = \sum_{i=1}^{N} \alpha_i \psi_i + \epsilon, \quad \psi_i \in \Psi.$$

$\epsilon$ is the residue. As $\Psi$ is over-complete, $\alpha_i, i = 1, 2..., N$ are called coefficient and usually $\alpha_i \neq < \mathbf{I}, \psi_i >$.

The image pyramids and wavelets are very successful in representing raw images in low level vision, but they are not very helpful in representing our perception for middle and high level vision tasks. For example, the wavelet decomposition does not correspond well with our perception of edges and object shapes. In the following we mention two main problems.

First, the frequency bases decomposition, nor the pyramid or the wavelet reconstruction, is not effective for representing image structures. Figures 4.(a) and (b) are respectively the

Gaussian and the Laplacian pyramids of a hexagon image. It is clearly that the boundary of the hexagon spreads across all levels of the Laplacian pyramid. Therefore, it consumes a large number of image bases to construct sharp edges. The reconstruction of the Gaussian pyramid by Gabor bases is shown in Figures 4.(c) and (d). The bases are computed by the matching pursuit algorithm[17]. Even with 500 bases, we still see blurry edges and aliasing effects.

Second, the image pyramid and wavelet representations are also ineffective for high entropy patterns, such as textures. A texture region often consumes a large number of image bases. However, in human perception, we are less sensitive to texture variations. A theoretical study on the coding efficiency is referred to in a companion paper[32].

These deficiencies suggest that we need to seek for a better model that (i) has a hyper-sparse dictionary to account for sharp edges, and (ii) separates textures from structures. This observation leads us to a primal sketch model.

### *B. Primal sketch representation*

A mathematical model of a primal sketch representation was proposed in [6], [8] to account for the generic and parsimonious token representation conjectured by Marr[18]. This representation overcomes the problems of the image pyramid and sparse coding mentioned above, and it bridges the image modeling between low and middle level vision. We use this representation to model perceptual transitions across scales.

Given an input image $\mathbf{I}$ on a lattice $\Lambda$, the primal sketch model divides the image into two parts: a "sketchable" part $\mathbf{I}_{\Lambda_{\rm sk}}$ for structures and a "non-sketchable" part $\mathbf{I}_{\Lambda_{\rm nsk}}$ for textures.

$$\mathbf{I} = (\mathbf{I}_{\Lambda_{\rm sk}}, \mathbf{I}_{\Lambda_{\rm nsk}}), \quad \Lambda = \Lambda_{\rm sk} \cup \Lambda_{\rm nsk}, \quad \Lambda_{\rm sk} \cap \Lambda_{\rm nsk} = \emptyset.$$

The structural part assumes an occlusion model, where $\Lambda_{\rm sk}$ is divided into a number of disjoint patches,

$$\Lambda_{\rm sk} = \cup_{k=1}^{N_{\rm sk}} \Lambda_{{\rm sk},k}, \quad \Lambda_{{\rm sk},k} \cap \Lambda_{{\rm sk},j} = \emptyset, \quad k \neq j.$$

Each image patch is selected from a primitive dictionary $\Delta$, and the residue follows *iid* Gaussian noise.

$$\mathbf{I}(u,v) = \mathbf{B}_k(u,v) + \epsilon(u,v),\ \epsilon(u,v) \sim N(0,\sigma_o), \quad \forall (u,v) \in \Lambda_{\text{sk},k},\ i = 1,...,N_{\text{sk}}.$$

where $k$ indexes the primitives in the dictionary $\Delta$ for translation $x, y$, rotation $\theta$, scaling $\sigma$, photometric contrast $\alpha$ and geometric warping $\vec{\beta}$,

$$k = (x_i, y_i, \theta_i, \sigma_i, \alpha_i, \vec{\beta}_i).$$

The primitive $B \in \Delta$ represents a step edge, a bar, an endpoint, a junction ("T" type or "Y" type), or a cross junction, etc. Figure 7 shows some examples of the primitives. Each primitive has $d+1$ points as landmarks shown on the top row of Figure 7. $d$ is the degree of connectivity, for example, a "T"-junction has $d = 3$, a bar has $d = 2$ and an endpoint has $d = 1$. The $(d+1)$'th point is the center of the primitive. When these primitives are aligned through their landmarks, we obtain a sketch graph $\mathbf{S}$, where each node is a primitive. A sketch graph is also called a Gestalt field and it follows a probability $p(\mathbf{S})$, which controls the graph complexity and favors good Gestalt organization, such as smooth connection between the primitives.

The remaining texture area $\Lambda_{\text{nsk}}$ is clustered into $N_{\text{nsk}} = 1 \sim 5$ homogeneous stochastic textures areas,

$$\Lambda_{\text{nsk}} = \cup_{j=1}^{N_{\text{nsk}}} \Lambda_{\text{nsk},j}.$$

Each follows a Markov random field model (FRAME)[36] with parameters $\eta_j$. These MRFs use the structural part $\mathbf{I}_{\Lambda_{\text{sk}}}$ as boundary condition.

$$\mathbf{I}_{\Lambda_{\text{nsk},j}} \sim p(\mathbf{I}_{\Lambda_{\text{nsk},j}} | \mathbf{I}_{\Lambda_{\text{sk}}}; \eta_j), j = 1,...,N_{\text{nsk}}.$$

The primal sketch is a two-level Markov model – one for the Gestalt field (object shapes) and the other for textures. For detailed description of the primal sketch model, please refer

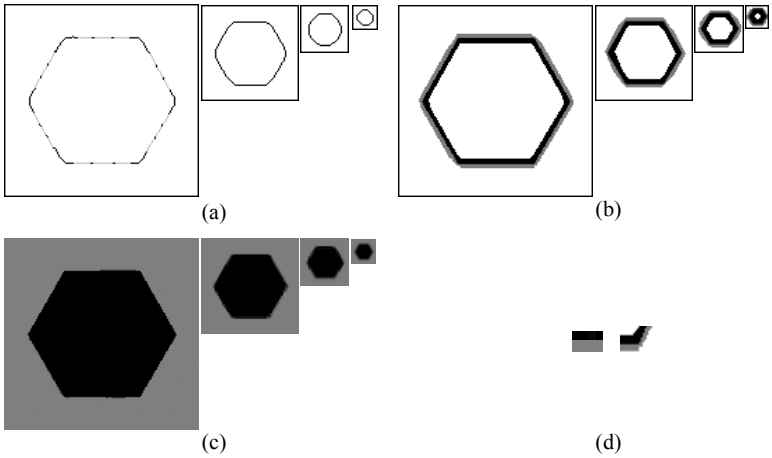



Fig. 5. Representing the hexagon image by primal sketch. (a) Sketch graphs at four levels as symbolic representation. They use two types of primitives shown in (d). The 6 corner primitives are shown in black, and step edges are in grey. The number of primitives used for the four levels are 24, 18, 12, 12, respectively (same as in Figure 4.c). (b) The sketchable part $\mathbf{I}_{\Lambda_{sk}}$ reconstructed by the two primitives with sketch $\mathbf{S}$. The remaining area (white) is non-sketchable (structureless): black or gray. (c) Reconstruction of the image pyramid after filling in the non-sketchable part. (d) Two primitives: a step-edge and an L-corner.

to [6], [7]. We show how the primal sketch represents the hexagon image over scales in Figure 5. It uses two types of primitives in Figure 5.(d). Only 24 primitives are needed at the highest resolution in Figure 5.(a) and the sketch graph is consistent over scales, although the number of primitives is reduced. As each primitive has sharp intensity contrast, there is no aliasing effects along the hexagon boundary in Figure 5.(c). The flat areas are filled in from the sketchable part in Figure 5.(b) through heat diffusion, which is a variation partial differential equation minimizing the Gibbs energy of a Markov random field. This is very much like image inpainting[2].

Compared to the image pyramids, the primal sketch has the following three characteristics.

1. The primitive dictionary is much sparser than the Gabor or Laplacian image bases, so that each pixel in $\Lambda_{sk}$ is represented by a single primitive. In contrast, it takes a few well-aligned image bases to represent the boundary.

2. In a sketch graph, the primitives are no longer independent but follow the Gestalt field,

so that the position, orientation, and intensity profile between adjacent primitives are regularized.

3. It represents the stochastic texture impression by Markov random fields instead of coding a texture in a pixel-wise fashion. The latter needs large scale image bases to code flat areas and still has aliasing effects.

## III. Perceptual uncertainty and transitions

In this section, we pose the perceptual transition problem in a Bayesian framework and attribute the transitions to the increase in the perceptual uncertainty of the posterior probability from fine to coarse in a Gaussian pyramid. Then, we identify three typical transitions over scales.

### *A. Perceptual transitions in image pyramids*

Visual perception is often formulated as Bayesian inference. The objective is to infer the underlying perceptual representation of the world denoted by $W$ from an observed image $\mathbf{I}$. Although it is a common practice in vision to compute the modes of a posterior probability as the most probable interpretations, we shall look at the entire posterior probability as the latter is a more comprehensive characterization of perception including uncertainty.

$$W \sim p(W|\mathbf{I};\Theta) \propto p(\mathbf{I}|W)p(W), \tag{1}$$

where $\Theta$ denotes the model parameters including a dictionary used in the generative likelihood. A natural choice for qualifying perceptual uncertainty is the entropy of the posterior distribution,

$$\mathcal{H}(p(W|\mathbf{I})) = -\sum_{W} p(W|\mathbf{I};\Theta)\log p(W|\mathbf{I};\Theta).$$

It is easy to show that when an image $\mathbf{I}$ is down-sampled to $\mathbf{I}_{\text{sm}}$ in a Gaussian pyramid, the uncertainty of $W$ will increase. This is expressed in a theorem below[7].

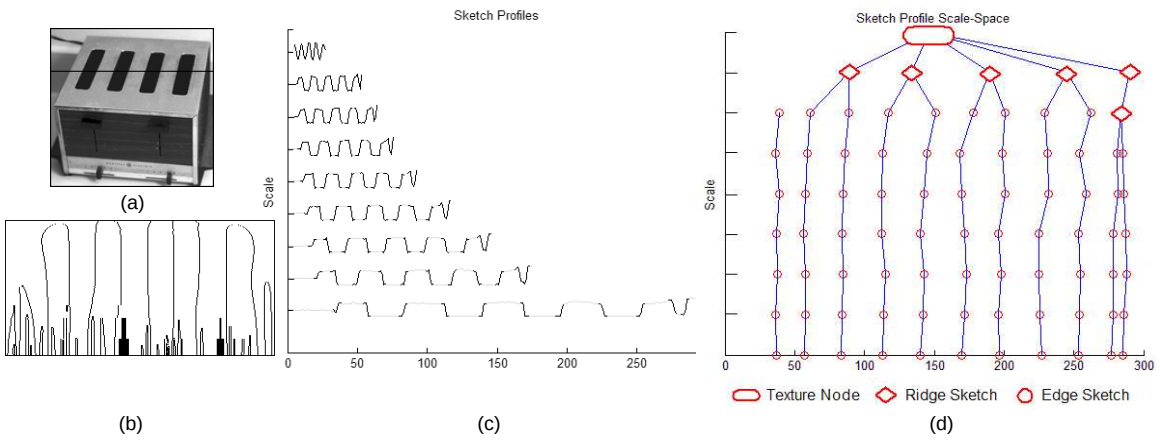



Fig. 6. Scale-space of a 1D signal. (a) A toaster image from which a line is taken as the 1D signal. (b) Trajectories of zero-crossings of the 2nd derivative of the 1D signal. The finest scale is at the bottom. (c) The 1D signal at different scales. The black segments on the curves correspond to primal sketch primitives (step edge or bar). (d) A symbolic representation of the sketch in scale-space with three types of transitions.

*Proposition 1:* Down-scaling increases the perceptual uncertainty,

$$\mathcal{H}(p(W|\mathbf{I}_{\rm sm});\Theta) \geq \mathcal{H}(p(W|\mathbf{I};\Theta)), \tag{2}$$

Consequently, we may have to drop some highly uncertain dimensions in $W$, and infer a reduced set of representation $W_{\rm sm}$ to keep the uncertainty of the posterior distribution of the pattern $p(W_{\rm sm}|\mathbf{I}_{\rm sm};\Theta_{\rm sm})$ at a reasonable small level. This corresponds to a model transition from $\Theta$ to $\Theta_{\rm sm}$, and a perceptual transition from $W$ to $W_{\rm sm}$. $W_{\rm sm}$ is of lower dimension than $W$.

$$\Theta \rightarrow \Theta_{\rm sm}, \quad W \rightarrow W_{\rm sm}.$$

For example, when we zoom out from the leaves or squares in Figures 1 and 2, some details of the leaves, such as the exact positions of the leaves, lose gradually, and we can no longer see the individual leaves. This corresponds to a reduced description of $W_{\rm sm}$.

### *B. Three types of transitions*

We identify three types of perceptual transitions in image scale-space. As they are reversible transitions, we discuss them either in down-scaling or up-scaling in a Gaussian

pyramid.

Following Witkin[31], we start with a 1D signal in Figure 6. The 1D signal is a horizontal slice from an image of a toaster in Figure 6.(a). Figure 6.(b) shows trajectories of zero-crossings of the 2nd derivative of the 1D signal. These zero-crossing trajectories are the signature of the signal in classical scale-space theory [31]. We reconstruct the signal using primal sketch with 1D primitives – step edges and ridges in Figure 6.(c) where the gray curves are the 1D signal at different scales and the dark segments correspond to the primitives. Viewing the signal from bottom-to-top, we can see that the "steps" are getting gentler; and at smaller scales, the pairs of steps merge into ridges. Figure 6.(d) shows a symbolic representation of the trajectories of the sketches tracked through scale-space and is called a sketch pyramid in 1D signal. Viewing the trajectories from top to bottom, we can see the perceptual transitions of the image from a texture pattern to several ridge type sketches, then split into number of step-edge type sketches when up-scaling. Figure 6.(d) is very similar to the zero-crossings in (b), except that this sketch pyramid is computed through probabilistic Bayesian inference while the zero-crossings are computed as deterministic features. The most obvious differences in this example are at the high resolutions where the zero-crossings are quite sensitive to small noise perturbations.

Now we show several examples for three types of transitions in images.

*Type 1: Blurring and sharpening of primitives.* Figure 7 shows some examples of image primitives in the dictionary $\Delta_{\text{sk}}$, such as step edges, ridges, corners, junctions. The top row shows the $d+1$ landmarks on each primitive with $d$ being the degree of connectivity. When an image is smoothed, the image primitives exhibit *continuous blurring* phenomenon shown in each column, or "sharpening" when we zoom-in. The primitives have parameters to specify the scale (blurring).

*Type 2: Mild jumps.* Figure 8 illustrates a perceptual scale-space for a cross with a four-level sketch pyramid $\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ and a series of graph grammar rules $\mathbf{R}_0, \mathbf{R}_1, \mathbf{R}_2$ for

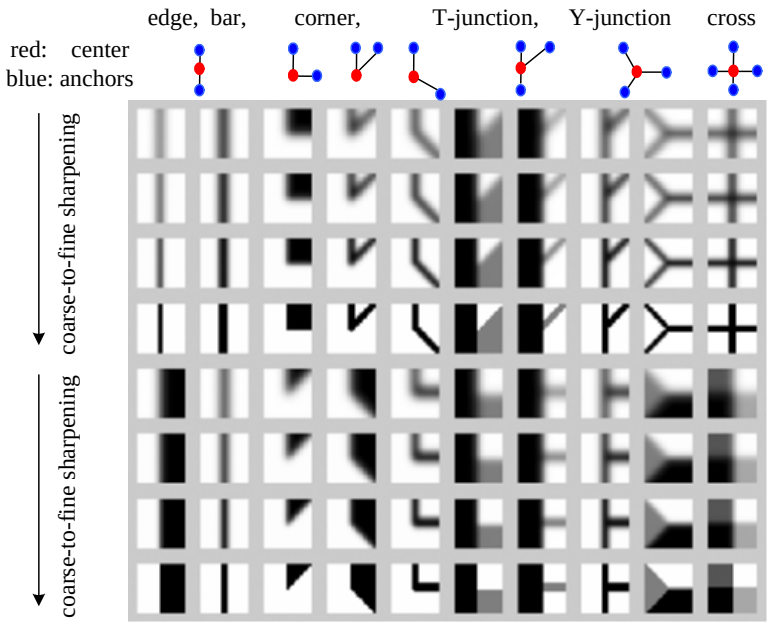



Fig. 7. Image primitives in the dictionary of primal sketch model are sharpened with four increasing resolutions from top to bottom. The blurring and sharpening effects are represented by scale parameters of the primitives.

graph contraction. Each $\mathbf{R}_k$ includes production rules $\gamma_{k,i}, i = 1, 2, ..., m(k)$ and each rule compresses a subgraph $g$ conditional on its neighborhood $\partial g$.

$$\mathbf{R}_k = \{\ \gamma_{k,i} : g_{k,i}|\partial g_{k,i} \rightarrow g'_{k,i}|\partial g_{k,i},\ i = 1, 2, ..., m(k)\ \}.$$

If a primitive disappear, then we have $g'_{k,i} = \emptyset$. Shortly, we shall show the 20 most frequent graph operators (rules) in Figure 10 in natural image scaling.

Graph contraction in a pyramid is realized by a series of rules,

$$\mathbf{S}_k \stackrel{\gamma_{k,1} \cdots \gamma_{k,m(k)}}{\longrightarrow} \mathbf{S}_{k+1}.$$

These operators explain the gradual loss of details (in red), for example, a cross shrinks to a dot, a pair of parallel lines merges into a bar (ridge), and so on. The operators are reversible depending on the upward or downward scaling.

*Type 3: Catastrophic texture-texton transition.* At certain critical scale, a large number of similar size primitives may disappear (or appear reversely) simultaneously. Figure 9 shows

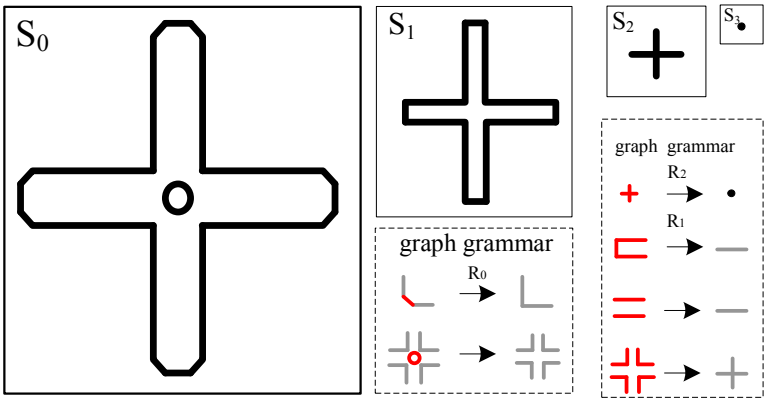



Fig. 8. An example of a 4-level sketch pyramid and corresponding graph operators for perceptual transitions.

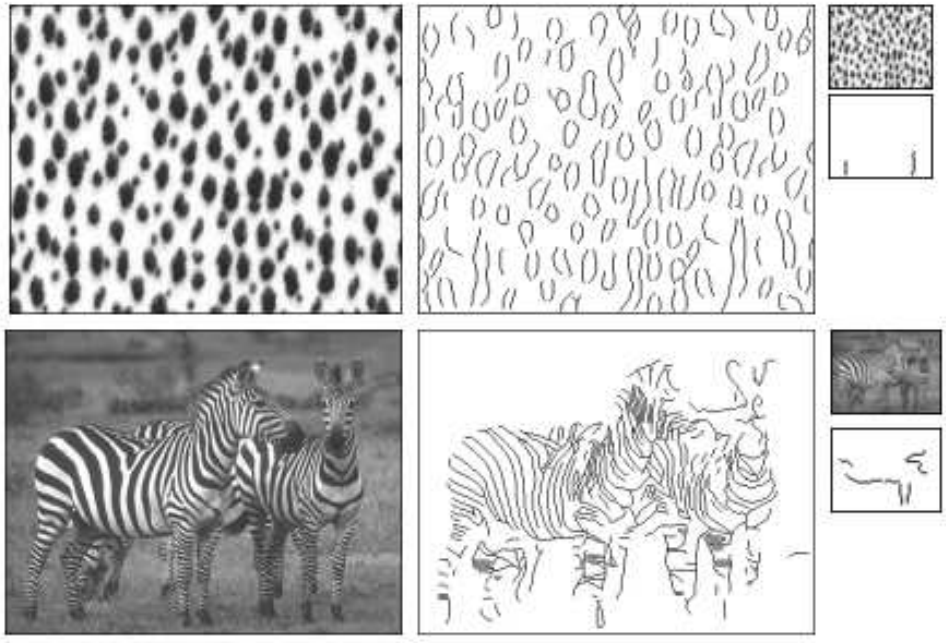


Fig. 9. Catastrophic texture-texton transition occurs when a large amount of primitives of similar sizes disappear (or appear) collectively.

two examples – cheetah dots and zebra stripe patterns. At high resolution, the edges and ridges for the zebra stripes and the cheetah blobs are visible, when the image is scaled down, we suddenly perceive only structureless textures. This transition corresponds to a significant model switching and we call it the catastrophic texture-texton transition. Another example is the leaves in Figure 1.

In summary, a sketch pyramid represents structural changes reversibly over scales which are associated with appearance changes. Each concept, such as a blob, a cross, a parallel bar only exists in a certain scale range (lifespan) in a sketch pyramid.

## IV. A generative model for perceptual scale-space

In this section, we formulate a generative model for the perceptual scale-space representation and pose the inference problem in a Bayesian framework.

We denote a Gaussian pyramid by $\mathbf{I}[0, n] = (\mathbf{I}_0, ..., \mathbf{I}_n)$. The perceptual scale-space representation, as shown in Figure 3, consists of two components – A sketch pyramid denoted by $\mathbf{S}[0, n] = (\mathbf{S}_0, ..., \mathbf{S}_n)$, and a series of graph grammar rules for perceptual transitions denoted by $\mathbf{R}[0, n-1] = (\mathbf{R}_0, \mathbf{R}_1, ..., \mathbf{R}_{n-1})$. Our objective is to define a joint probability $p(\mathbf{I}[0, n], \mathbf{S}[0, n], \mathbf{R}[0, n-1])$ so that an optimal sketch pyramid and perceptual transitions can be computed though maximizing the joint posterior probability $p(\mathbf{S}[0, n], \mathbf{R}[0, n-1]|\mathbf{I}[0, n])$. This is different from computing each sketch level $\mathbf{S}_k$ from $\mathbf{I}_k$ independently. The latter may cause "flickering" effects such as the disappearance and reappearance of an image feature across scales.

### *A. Formulation of a single level primal sketch*

Following the discussion in Section (II-B), the generative model for primal sketch is a joint probability of a sketch $\mathbf{S}$ and an image $\mathbf{I}$,

$$p(\mathbf{I}, \mathbf{S}; \Delta_{\text{sk}}) = p(\mathbf{I}|\mathbf{S}; \Delta_{\text{sk}})p(\mathbf{S}).$$

The likelihood is divided into a number of primitives and textures,

$$p(\mathbf{I}|\mathbf{S}; \Delta_{\text{sk}}) \propto \prod_{k=1}^{N_{\text{sk}}} \exp\{- \sum_{(u,v)\in\Lambda_{\text{sk},k}} (\frac{\mathbf{I}(u, v) - B_k(u, v))^2}{2\sigma_o^2}\} \cdot \prod_{j=1}^{N_{\text{nsk}}} p(\mathbf{I}_{\Lambda_{\text{nsk},j}} \,|\, \mathbf{I}_{\Lambda_{\text{sk}}}; \eta_j).$$

$\mathbf{S} =< V, E >$ is an attribute graph. $V$ is a set of primitives in $\mathbf{S}$

$$V = \{B_k, k = 1, 2, ..., N_{\text{sk}}\}.$$

$E$ denotes the connectivity for neighboring structures,

$$E = \{e =< i, j >: B_i, B_j \in V\}$$

The prior model $p(\mathbf{S})$ is an inhomogeneous Gibbs model defined on the attribute graph to enforce some Gestalt properties, such smoothness, continuity and canonical junctions:

$$p(\mathbf{S}) \propto \exp\{-\sum_{d=0}^{4} \zeta_d N_d - \sum_{<i,j>\in E} \psi(B_i, B_j)\},$$

where the $N_d$ is the number of primitives in $\mathbf{S}$ whose degree of connectivity is $d$. $\zeta_d$ is the parameter that controls the number of primitives $N_d$ and thus the density. In our experiments, we choose $\zeta_0 = 1.0$, $\zeta_1 = 5.0$, $\zeta_2 = 2.0$, $\zeta_3 = 3.0$, $\zeta_4 = 4.0$. The reason we give more penalty for terminators is that the Gestalt laws favor closure and continuity properties in perceptual organization. $\psi(B_i, B_j)$ is a potential function of the relationship between two vertices, e.g. smoothness and proximity. A detailed description is referred to [7].

*B. Formulation of a sketch pyramid*

Because of the intrinsic perceptual uncertainty in the posterior probability (Eq.1), the sketch pyramid $\mathbf{S}_k, k = 0, 1, ...n$ will be inconsistent if each level is computed independently. For example, we may observe a "flickering" effect when we view the sketches from coarse-to-fine (see Figures 16.b).

To ensure monotonic graph transitions and consistency of a sketch pyramid, we define a common set of graph operators

$$\Sigma_{\text{gram}} = \{\mathcal{T}_{\emptyset}, \mathcal{T}_{dn}, \mathcal{T}_{me2r}, ...\}.$$

They stand, respectively, for null operation (no topology change), death of a node, merging a pair of step-edges into a ridge, etc. Figure 10 shows a graphical illustration of twenty down-scale graph operators (rules), which are identified and learned through a supervised learning procedure in Section V. Figure 11 shows a few examples in images by rectangles where the operators occur.

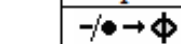
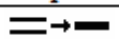
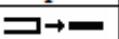
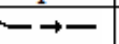
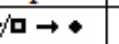
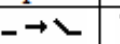
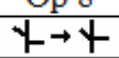
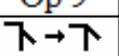
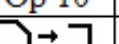
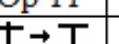
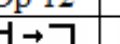
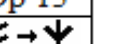
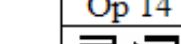
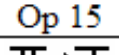
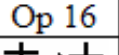
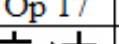
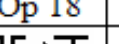
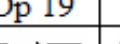
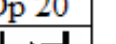
| Op 0 | Op 1 | Op 2 | Op 3 | Op 4 | Op 5 | Op 6 |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
| Op 7 | Op 8 | Op 9 | Op 10 | Op 11 | Op 12 | Op 13 |
|  |  |  |  |  |  |  |
| Op 14 | Op 15 | Op 16 | Op 17 | Op 18 | Op 19 | Op 20 |
|  |  |  |  |  |  |  |

Fig. 10. Twenty graph operators identified from down-scaling image pyramids. For each operator, the left graph turns into the right graph when the image is down-scaled.

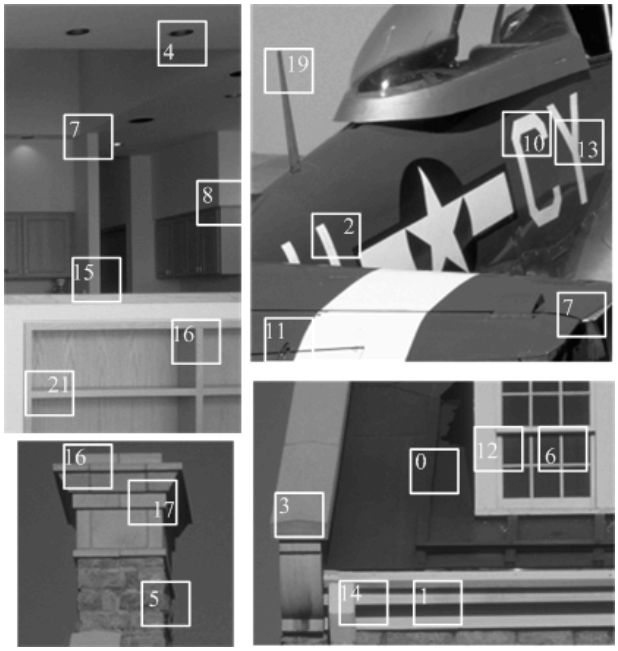



Fig. 11. Each square marks the image patch where the graph operator occurs, and each number in the squares correspond to the index of graph operators listed in Fig.10.

Each rule $\gamma_n \in \Sigma_{\text{gram}}$ is applied to a subgraph $g_n$ with neighborhood $\partial g_n$ and replaces it by a new subgraph $g'_n$. The latter has smaller size for monotonicity, following the Proposition 1.

$$\gamma_n : g_n|\partial g_n \rightarrow g'_n|\partial g_n, \quad |g'_n| \leq |g_n|.$$

Each rule is associated with a probability depending on its attributes,

$$\gamma_n \sim p(\gamma_n) = p(g_n \rightarrow g'_n|\partial g_n), \quad \gamma_n \in \Sigma_{\text{gram}}.$$

$p(\gamma_n)$ will be decided through supervised learning described in Section (V).

As discussed previously, transitions from $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$ are realized by a sequence of $m(k)$ production rules $\mathbf{R}_k$,

$$\mathbf{R}_k = (\gamma_{k,1}, \gamma_{k,2}, ..., \gamma_{k,m(k)}),\ \gamma_{k,i} \in \Sigma_{\text{gram}}.$$

The order of the rules matters and the rules constitute a path in the space of sketch graphs from $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$.

The probability for the transitions from $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$ is,

$$p(\mathbf{R}_k) = p(\mathbf{S}_{k+1} \,|\, \mathbf{S}_k) = \prod_{i=1}^{m(k)} p(\gamma_{k,i}).$$

A joint probability of the scale-space is

$$p(\mathbf{I}[0,n], \mathbf{S}[0,n], \mathbf{R}[0,n-1]) = \prod_{k=0}^{n} p(\mathbf{I}_k|\mathbf{S}_k; \Delta_{\text{sk}}) \cdot p(\mathbf{S}_0) \cdot \prod_{k=0}^{n-1} \prod_{j=1}^{m(k)} p(\gamma_{k,j}), \tag{3}$$

*C. A criterion for perceptual transitions*

A central issue for computing a sketch pyramid and associated perceptual transitions is to decide which structure should appear at which scale. In this subsection, we shall study a criterion for the transitions. This is posed as a model comparison problem in the Bayesian framework.

By induction, suppose $\mathbf{S}$ is the optimal sketch from $\mathbf{I}$. At the next level, image $\mathbf{I}_{\text{sm}}$ has decreased resolution, and so $\mathbf{S}_{\text{sm}}$ has less complexity following Proposition 1. Without loss of generality, we assume that $\mathbf{S}_{\text{sm}}$ is reduced from $\mathbf{S}$ by a single operator $\gamma$.

$$\mathbf{S} \stackrel{\gamma}{\Longrightarrow} \mathbf{S}_{\text{sm}}.$$

we compute the ratio of the posterior probabilities.

$$\begin{aligned} \delta(\gamma) &\triangleq \log \frac{p(\mathbf{I}_{\text{sm}}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}_{\text{sm}}|\mathbf{S})} + \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})} && (4) \\ &= \log \frac{p(\mathbf{S}_{\text{sm}}|\mathbf{I}_{\text{sm}})}{p(\mathbf{S}|\mathbf{I}_{\text{sm}})}, \quad \text{if } \lambda_\gamma = 1. && (5) \end{aligned}$$

The first log-likelihood ratio term is usually negative even for a good choice of $\mathbf{S}_{\text{sm}}$, because a reduced generative model will not fit an image as well as the complex model $\mathbf{S}$. However, the prior term $\log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})}$ is always positive to encourage simpler models.

Intuitively, the parameter $\lambda_\gamma$ balances the model fitting and the model complexity. As we know in Bayesian decision theory, a decision may not be only decided by the posterior probability or coding length, it is also affected by some cost function (not simply $0-1$ loss) related to perception. The cost function is summarized into $\lambda_\gamma$ for each $\gamma \in \Sigma_{\text{gram}}$.

- $\lambda_\gamma = 1$ corresponds to the Bayesian (MAP) formulation, with $0-1$ loss function.
- $\lambda_\gamma > 1$ favors applying the operator $\gamma$ earlier in the down-scaling process, and thus the simple description $\mathbf{S}_{\text{sm}}$.
- $\lambda_\gamma < 1$ encourages "hallucinating" features when they are unclear.

Therefore, $\gamma$ is accepted, if $\delta(\gamma) > 0$. More concretely, a graph operator $\gamma$ occurs if

$$\log \frac{p(\mathbf{I}_{\text{sm}}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}_{\text{sm}}|\mathbf{S})} + \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})} > 0 \ \text{ and } \ \log \frac{p(\mathbf{I}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}|\mathbf{S})} + \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})} < 0. \tag{6}$$

The transitions $\mathbf{R}_k$ between $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$ consist of a sequence of such greedy tests. In the next section, we learn the range of the parameters $\lambda_\gamma$ for each $\gamma \in \Sigma_{\text{gram}}$ from human experiments.

## V. Supervised learning of parameters

In this section, we learn a set of the most frequent graph operators $\Sigma_{\text{gram}}$ in image scaling and learn a range of parameter $\lambda_\gamma$ for each operator $\gamma \in \Sigma_{\text{gram}}$ through simple human experiments. The learning results will be used to infer sketch pyramids in Section VI.

We selected 50 images from the Corel image database. The content of these images covers a wide scope: natural scenes, architectures, animals, human beings, and man-made objects. Seven graduate students with and without computer vision background were selected randomly from different departments as subjects. We provided a computer graphics user interface (GUI) for the 7 subjects to identify and label graph transitions in the 50 image

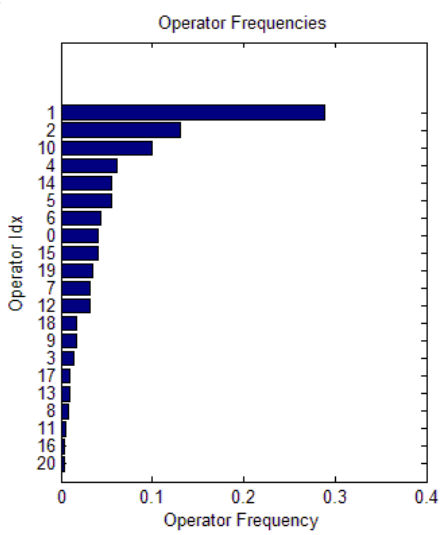



Fig. 12. The frequencies of the top 20 graph operators shown in Fig.10.

pyramids. The labeling procedure is as follows. First, the software will load a selected image $\mathbf{I}_0$, and build a Gaussian pyramid $(\mathbf{I}_0, \mathbf{I}_2, ..., \mathbf{I}_n)$. Then the sketch pursuit algorithm [6] is run on the highest resolution to extract a primal sketch graph, which is then manually edited to fix some errors to get a perfect sketch $\mathbf{S}_0$.

Next, the software builds a sketch pyramid upwards by generating sketches by shrinking the sketch proportionally to the next gaussian level (starting with $\mathbf{S}_0$) one by one until the coarsest scale. The subjects will search across both the Gaussian and sketch pyramids to label places where they think graph editing is needed, e.g. some sketch disappears, or a pair of double edge sketches may be replaced by a ridge sketch. Each type of transition corresponds to a graph operator or a graph grammar rule. All these labeled transitions are automatically saved across all scale and for the 50 images.

The following are some results from this process.

*Learning result 1: frequencies of graph operators.* Figure 10 shows the top 20 graph operators which have been applied most frequently in the 50 images among the 7 subjects. Figure 12 plots the relative frequency for the 20 operators. It is clear that the majority of perceptual transitions correspond to operator 1, 2, and 10, as they are applied to the most

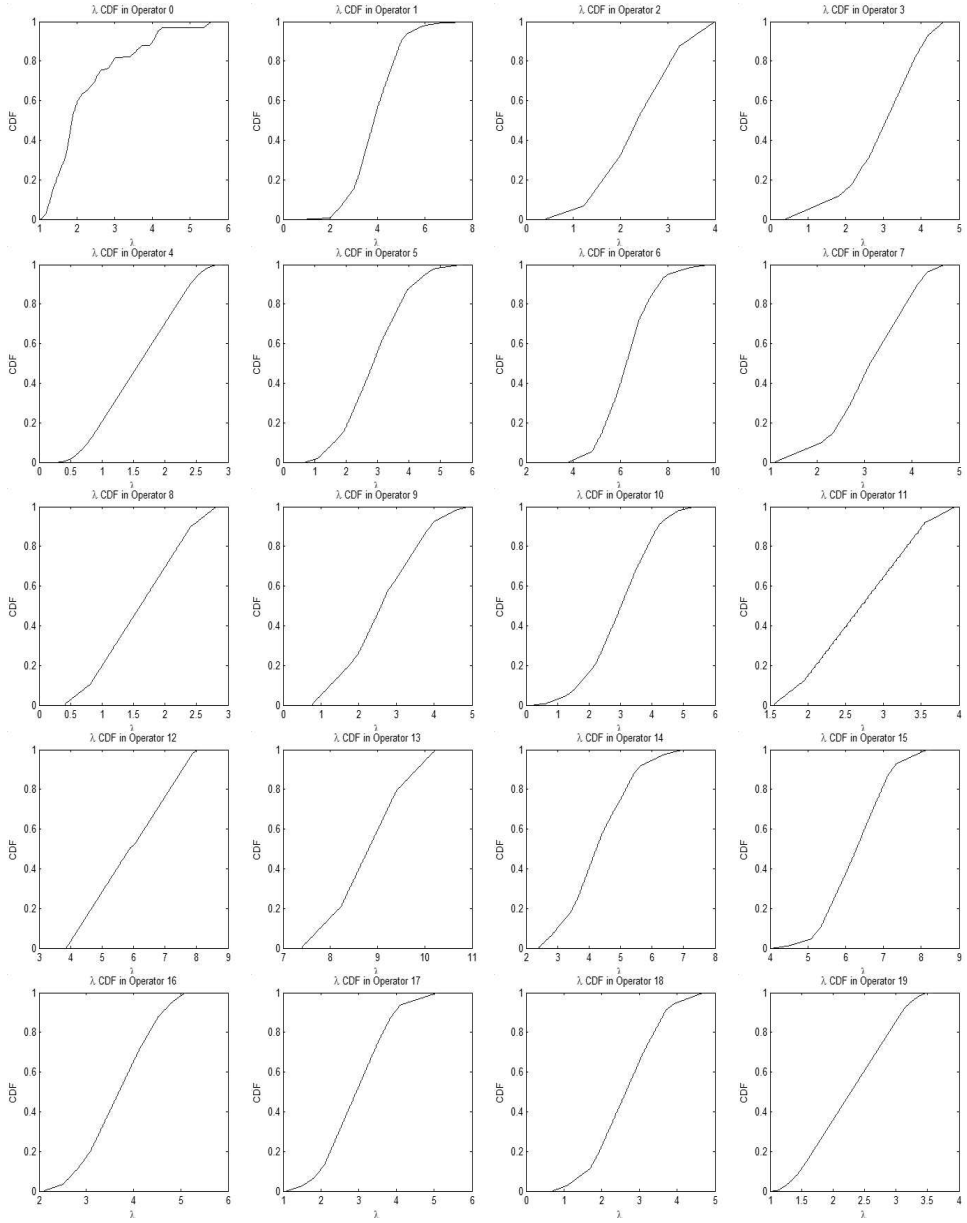
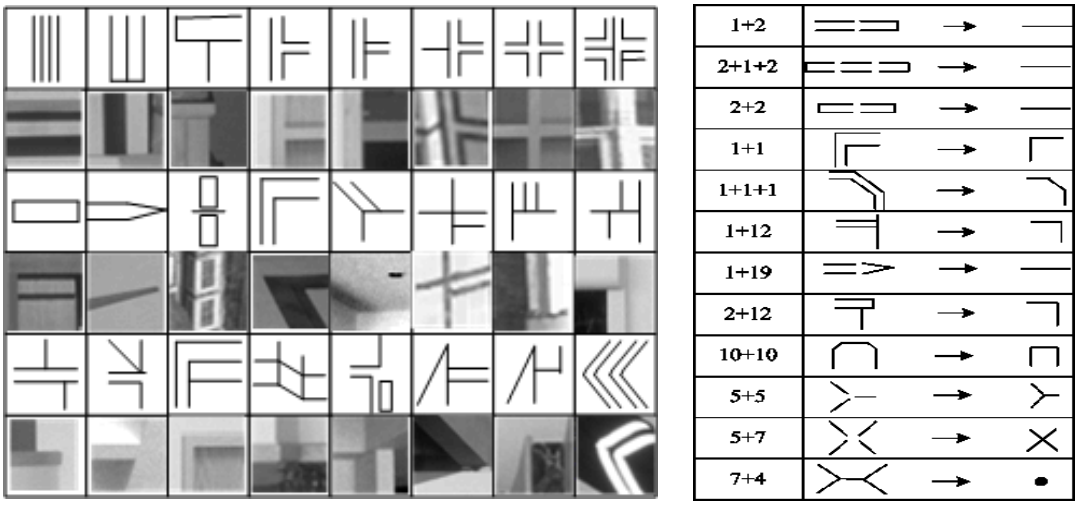



Fig. 13. CDFs of $\lambda$ for each graph grammar rule or operator listed in Fig.10.

frequently observed and generic structures in images.

*Learning result 2: range of $\lambda_\gamma$.* The exact scale where a graph operator $\gamma$ is applied varies among the human subjects. Suppose an operator $\gamma$ occurs between scales $\mathbf{I}$ and $\mathbf{I}_{\text{sm}}$, then we can compute the ratios.

$$a_1 = -\log \frac{p(\mathbf{I}_{\text{sm}}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}_{\text{sm}}|\mathbf{S})},\ b_1 = \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})},\ a_2 = -\log \frac{p(\mathbf{I}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}|\mathbf{S})},\ b_2 = \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})}.$$




a) (b)

Fig. 14. (a) Frequently observed local image structures. The line drawings above each image patch are the corresponding subgraphs in its sketch graph. (b) The most frequently co-occurring operator sets. The numbers in the left column are the indices to the top 20 operators listed in Figure 10. The diagrams in the right column are the corresponding subgraphs and transition examples.

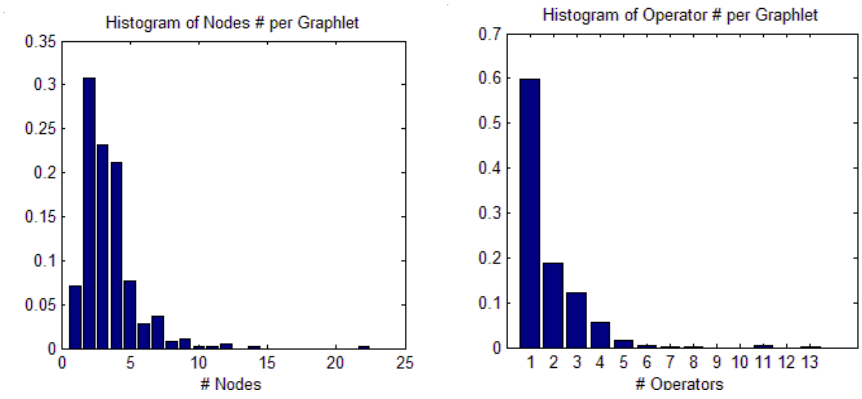
By the inequalities in Eq.6, we can determine an interval for $\lambda_\gamma$

$$\frac{a_1}{b_1} < \lambda_\gamma < \frac{a_2}{b_2}.$$

The interval above is for a specific occurrence of $\gamma$ and it is caused by finite levels of a Gaussian pyramid. By accumulating the intervals for all instances of the operator $\gamma$ in the 50 image and the 7 subjects, we obtain a probability (histogram) for $\gamma$. Figure 13 shows the cumulative distribution functions (CDF) of $\lambda_\gamma$ for the top 20 operators listed in Figure 10.

*Learning result 3: Graphlets and composite graph operators.* Often, several graph operators occur simultaneously at the same scale in a local image structure or subgraph of a primal sketch. Figure 14.(a) shows some typical subgraphs where multiple operators happen frequently. We call these subgraphs "graphlets". By using the "*Apriori Algorithm*"[1], we find the most frequently associated graph operators in Figure 14.(b). We call these *composite operators*.

Figure 15 shows the frequency counts for the graphlets and composite graph operators.




(a) (b)

Fig. 15. Frequency of the graphlets and composite operators. (a) Histogram of the number of nodes per "graphlet" where perceptual transitions happen simultaneously. (b) Histogram of the number of operators applied to each "graphlet" .

Figure 15.(a) shows that a majority of perceptual transitions involves sub-graphs with no more than 5 primitives (nodes). Figure 15.(b) shows that the frequency for the number of graph operators involved in each composite operator. We include the single operator as a special case for comparison of frequency.

In summary, the human experiments on the sketch pyramids set the parameters $\lambda_\gamma$, which will decide the threshold of transitions. In our experiments, it is evident that human vision has two preferences in comparison with the pure maximum posterior probability (or MDL) criterion (i.e. $\lambda_\gamma = 1,\ \forall\gamma$.)

- Human vision has a strong preference for simplified descriptions. As we can see that in Figure 13, $\lambda_\gamma > 1$ for most operators. Especially, if there are complicated structures, human vision is likely to simplify the sketches. For example, $\lambda_\gamma$ goes to the range of $[4, 9]$ for operators No. 13, 14, 15.
- Human vision may hallucinate some features, and delay their disappearance, for example, $\lambda_\gamma < 1$ for operator No.4.

These observations become evident in our experiments in the next section.

## VI. Upwards-downwards inference and experiments

In this section, we briefly introduce an algorithm that infers hidden sketch graphs $\mathbf{S}[0, n]$ upwards and downwards across scales using the learned models $p(\lambda_\gamma)$ for each grammar rule. Then we show experiments of computing sketch pyramids.

### A. The inference algorithm

Our goal is to infer consistent sketch pyramids from Gaussian image pyramids, together with the optimal path of transitions by maximizing a Bayesian posterior probability,

$$\begin{aligned} & (\mathbf{S}[0, n], \mathbf{R}[0, n-1])^* \\ = \ & \arg\max p(\mathbf{S}[0, n], \mathbf{R}[0, n-1] | \mathbf{I}[0, n]) \\ = \ & \arg\max \prod_{k=0}^{n} p(\mathbf{I}_k | \mathbf{S}_k; \Delta_k) \cdot p(\mathbf{S}_0) \prod_{k=1}^{n} \prod_{j=1}^{m(k)} p(\gamma_{k,j}). \end{aligned}$$

Our inference algorithm consists of three stages.

*Stage I: Independent sketching.* We first apply the primal sketch algorithm[6], [7] to image $\mathbf{I}_0$ at the bottom of a Gaussian pyramid to compute $\mathbf{S}_0$. Then we compute $\mathbf{S}_k$ from $\mathbf{I}_k$ using $\mathbf{S}_{k-1}$ as initialization for $k = 1, 2, ..., n$. As each level of sketch is computed independently by MAP estimation, the consistency of the sketch pyramid is not guaranteed. Figure 16.(b) shows the sketch pyramid where we observe some inconsistencies in the sketch graph across scales.

*Step II: Bottom-up graph matching.* This step gives the initial solution of matching sketch graph $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$. We adopt standard graph matching algorithm discussed in [35], [11] and [29] to match attribute sketch graphs across scales. Here, we briefly report how we compute the matching with a bottom-up approach.

A sketch as an attribute node in a sketch graph (as shown in Figure 16) has the following properties.

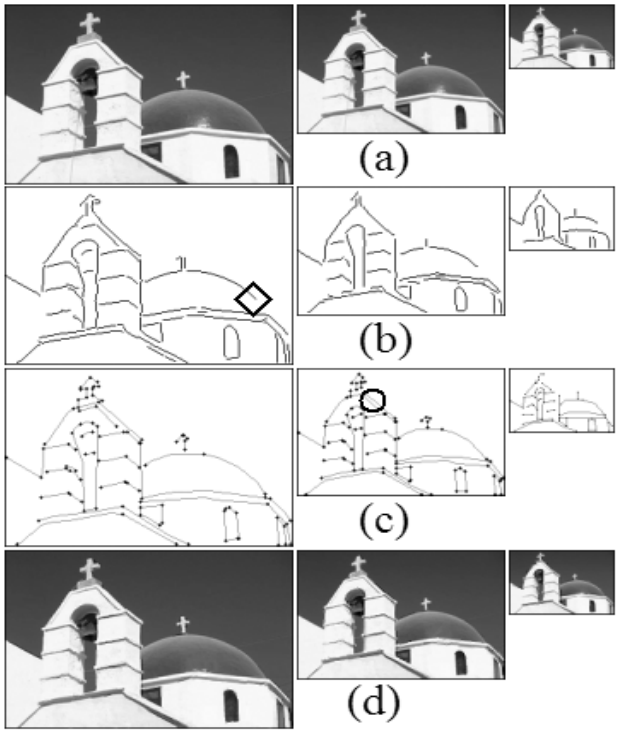



Fig. 16. A church image in scale-space. (a) Original images across scales. The largest image size is $241 \times 261$. (b) Initial sketches computed independently at each level by algorithm. (c) Improved sketches across scales. The dark dots indicate end points, corners and junctions. d) Synthesized images by the sketches in (c). The symbols mark the perceptual transitions.

1. Normalized length $\mathbf{l}$ by its scale.
2. Shape $\mathbf{s}$. A set of control points connectively define the shape of a sketch.
3. Appearance $\mathbf{a}$. (Pixel intensities of a sketch.)
4. Degree $\mathbf{d}$. (Number of connections at each end of a sketch. )

In the following, we also use $\mathbf{l}, \mathbf{s}, \mathbf{a}, \mathbf{d}$ as functions on sketch node $i$, i.e., each returns the corresponding feature. For example, $\mathbf{l}(i)$ tells the normalized length of the $i$'th sketch node by its scale.

A match between the $i$'th sketch node at scale $k$ (denoted as $v_i$) and the $j$'th sketch node at scale $k+1$ (denoted as $v_j$) is defined as a probability:

$$P_{match}[v_i, v_j] = \frac{1}{Z} \exp\{-\frac{(\mathbf{l}(i) - \mathbf{l}(j))^2}{2\sigma_c^2} - \frac{(\mathbf{s}(i) - \mathbf{s}(j))^2}{2\sigma_s^2} - \frac{(\mathbf{a}(i) - \mathbf{a}(j))^2}{2\sigma_a^2} - \frac{(\mathbf{d}(i) - \mathbf{d}(j))^2}{2\sigma_d^2}\}$$

where $\sigma_{\cdot}$'s are the variances of the corresponding features. This similarity measurement is also used in the following graph editing part to compute the system energy. When matching $\mathbf{S}_k = (v_i(k), i = 1, \ldots, n)$ and $\mathbf{S}_{k+1} = (v_i(k+1), i = 1, \ldots, n)$, where $n$ is the larger number of sketches in either of the two graphs, it is reasonable to allow some sketches in $\mathbf{S}_k$ to map to null, or multiple sketches in $\mathbf{S}_k$ map to the same sketch in $\mathbf{S}_{k+1}$, and vice versa. Thus, the similarity between graph $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$ is defined as a probability:

$$P[\mathbf{S}_k, \mathbf{S}_{k+1}] = \prod_{i=1}^{n} P_{match}[v_i(k), v_i(k+1)]$$

The graph matching results are used as an initial match to feed into the following MCMC process.

*Step III: Matching and editing graph structures by MCMC sampling.* Because of the intrinsic perceptual uncertainty in the posterior probability, and the huge and complicated solution space for hidden dynamic graph structures, we adopt the MCMC reversible jumps [5] to match and edit iteratively the computed sketch graphs both upwards and downwards in scale-space, so as to infer the hidden dynamic graph structures and to find the optimal transition paths.

Our Markov chain consists of twenty pairs of reversible jumps (listed in Figure 10) to adjust the matching of adjacent graphs in a sketch pyramid based on the initial matching results in Step II, so as to achieve a high posterior probability. These reversible jumps correspond to the grammar rules in $\Sigma_{gram}$. Each pair is selected probabilistically and they observe the detailed balance equations. Each move in the Markov chain design is a reversible jump between two states $A$ and $B$ realized by a Metropolis-Hastings method [19]. We design a pair of proposal probabilities for moving from $A$ to $B$, with $q(A \to dB) = q(B|A)dB$, and back with $q(B \to dA) = q(A|B)dA$. The proposed move is accepted with probability

$$\alpha(A \to B) = \min(1, \frac{q(A|B)dA \cdot p(B|\mathbf{I}^{\text{obs}}[1,\tau])dB}{q(B|A)dB \cdot p(A|\mathbf{I}^{\text{obs}}[1,\tau])dA}).$$

These MCMC moves simulate a Markov chain with invariant probability $p(\mathbf{S}[0,n], \mathbf{R}[0,n-$

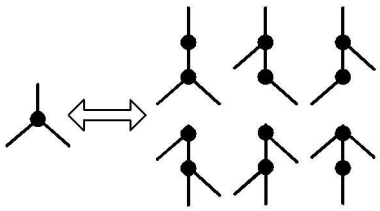


Fig. 17. Split/merge graph operation diagram. A vertex can be split into two vertices with one of six edge configurations.

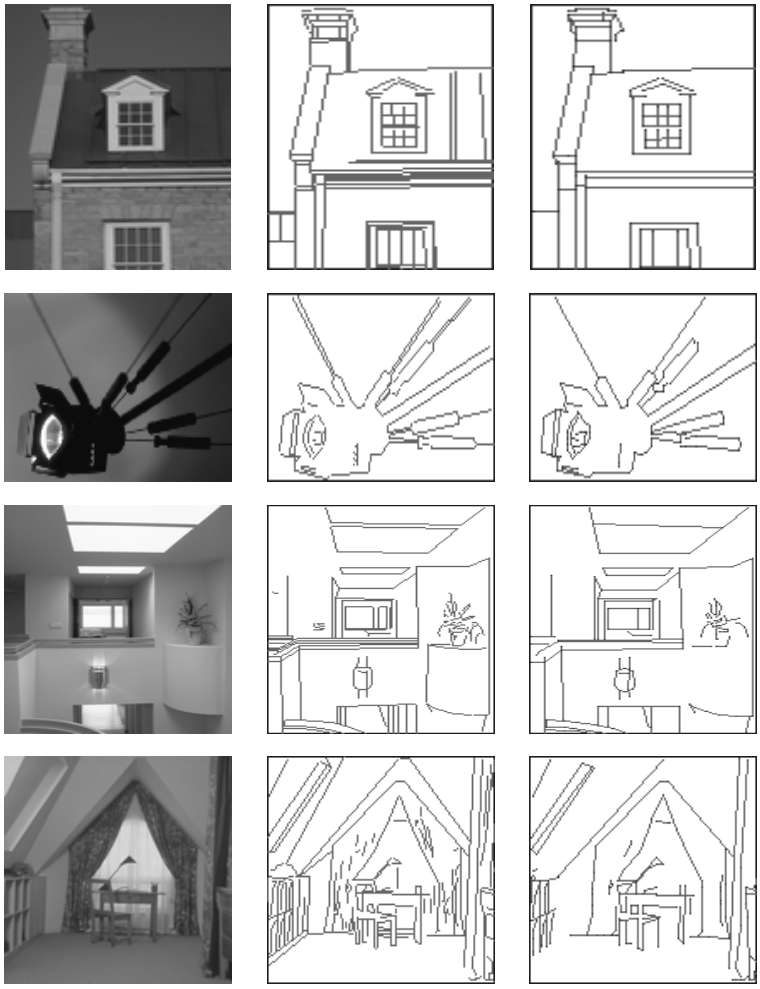
1] $|\mathbf{I}[0,n])$. In the previous sections, we have discussed each probability model in Eq.3, including the image photometric model $p(\mathbf{I}_k|\mathbf{S}_k)$, the primal sketch geometric model $p(\mathbf{S}_k)$ and the graph grammar rule model $p(\gamma_i)$. These probability models are used in this inference process when sampling from the posterior probability. The design of reversible jumps is very similar to the design in [29], [7]. Due to the page limit, we only introduce one pair of Markov chain moves – split/merge (Operator 7 in Figure 10). The moves are illustrated in Figure 17 and they are jump processes between two states $A$ and $B$, where

$$\begin{aligned} A &= (n, \mathbf{S} =< (V_-, v_j), (E_-, e_{i,j}) >) \\ &\rightleftharpoons (n-1, \mathbf{S}' =< V_-, E_- >) = B, \end{aligned}$$

where $n$ is the number of sketches in sketch graph $\mathbf{S}$. $V_-$ and $E_-$ denote the unchanged sketch set and edge set, respectively. $e_{i,j}$ is the edge between sketches $v_i$ and $v_j$, and $v_j$ is the sketch disappeared after merging. We define the proposal probabilities as follows.

$$\begin{aligned} q(A \to B) &= q_{s/m} \cdot q_m \cdot q(i) \cdot q(j) \\ q(B \to A) &= q_{s/m} \cdot q_s \cdot q'(i) \cdot q(pattern). \end{aligned}$$

$q_{s/m}$ is the probability for selecting this split/merge move among all possible graph operations. $q_m$ and $q_s$ is the probability to choose either split or merge, respectively, where $q_m + q_s = 1$. $q(i)$ is the probability of selecting $v_i$ as the anchor vertex for the other vertex to merge into, which is usually set to $1/n$. $q(j)$ is the probability to choose $v_j$ from $v_i$'s neighbors, which is set to be inversely proportional to the distance between $v_i$ and $v_j$. When



(a) (b) (c)

Fig. 18. Sketch graph comparison between applying simple death operator and applying learned perceptual graph operators. (a)Original images. (b)Sketch graphs at scale level 7 with only death operator applied. (c) Sketch graphs at scale level 7 with learned perceptual graph operator applied.

proposing a split move, $q'(i)$ is the probability to choose $v_i$. It is assumed to be uniform among those qualified vertices. When a sketch with $m$ edges is split, there are $1/(2^m - 2)$ ways for two vertices to share these $m$ edges. Therefore, $q(pattern)$ is set to be $1/(2^m - 2)$.

### *B. Experiments on sketch pyramids*

We successfully applied the inference algorithm on 20 images to compute their sketch pyramids in perceptual scale-space. In this subsection, we show some of them to illustrate

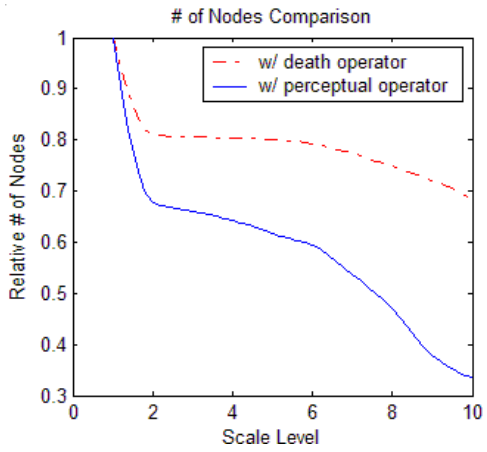



Fig. 19. A comparison of relative node number in sketch graphs at each scale between applying simple death operators and applying the learned perceptual graph operators. Images at scale level 1 has the highest resolution, containing maximum number of nodes in the sketch graphs, which is normalized to 1.0. The dashed line shows the number of sketches at each scale level when simple death operator is applied. The solid line shows the number of sketches after applying perceptual operators learned from human subjects.

the inference results.

*Inference result 1: Sketch consistency.* Figures 16.(c) shows examples of the inferred sketch pyramids obtained by the MCMC sampling with the learned graph grammar rules. We compare the results with the initial bottom-up sketches (b) where each level is computed independently and in a greedy way. The improved results show consistent graph matching across scales.

*Inference result 2: Compactness of perceptual sketch pyramid.* In Figure 18, we compare the inferred sketch graphs in column (c) with the sketch graph obtained by only applying death operator in column (b). We can see that the sketch graphs inferred from perceptual scale-space are more compact and closer to human perception.

Figure 19 compares the compactness of the sketch pyramid representation obtained by applying learned perceptual graph operators against that rendered by applying only simple death operators. The compactness is measured by the number of sketches. Images at scale level 1 has the highest resolution, containing maximum number of sketches in the sketch

graph, and the number of sketches at this scale level is normalized to 1.0. When scaling down, the sketch nodes are getting shorter and shorter, till they finally disappear, where the death operator applies. Thus the higher the scale level, the fewer the sketch nodes. The dashed line shows the relative number of sketches at each scale level with only the simple death operator applied when scaling down. The solid line shows the relative number of sketches at each scale level by applying perceptual operators learned from human subjects. The number of sketches indicates the coding length of sketch graphs. As human beings always prefer simpler models without perceptual loss, the sketch pyramid inferred with learned perceptual graph operators is a more compact representation.

In summary, from the above inference results, we conclude that the inferred perceptual sketch pyramid is a very good representation for human perception. By properly addressing the perceptual transition issue in perceptual scale-space, we expect performance improvements in many vision applications. For example, by explicitly modeling these quantum jumps, we can finally begin to tackle visual scaling tasks that must deal with objects and features that look fundamentally different across scales. In object recognition, for example, the features that we use to recognize an object from far away may belong to a different class than those we use to recognize it at a closer distance. The perceptual scale-space will allow us to seamlessly connect these scale-variant features in a probabilistic chain. In the following section, we show an application based on an inferred perceptual sketch pyramid.

## VII. An application – adaptive image display

Because of the improving resolution of digital imaging and use of portable devices, recently there have been emerging needs , as stated in [33], for displaying large digital images (say $\Lambda = 2048 \times 2048$ pixels) on small screens (say $\Lambda_o = 128 \times 128$ pixels), such as in PDAs, cellular phones, image icon display on PCs, and digital cameras. To reduce manual operations involved in browsing within a small window, it is desirable to display a short movie for a

Fig. 20. A tour over a Gaussian pyramid. Visiting the decomposed quad-tree nodes in a sketch pyramid is an efficient way to automatically convey a large image's informational content.

"tour" that the small window $\Lambda_o$ flying through the lattice $\Lambda$, so that most of the *information* in the image is displayed in as few frames as possible. These frames are snapshots of a Gaussian pyramid at different resolutions and regions. For example, one may want to see a global picture at a coarse resolution and then zoom in some interesting objects to view the details.

Figure 20 shows a demo on a small PDA screen. The PDA displays the original image at the upper-left corner, within which a window (white) indicates the area and resolution shown on the screen. To do so, we decompose a Gaussian pyramid into a quadtree representation shown in Figure 21. Each node in the quadtree hierarchy represent a square region of a constant size. We say a quadtree node is "visited" if we show its corresponding region on the screen. Our objective is to design a visiting order of some nodes in the quadtree. The quadtree simplifies the representation but causes border artifacts when an interesting object is in the middle and has to be divided into several nodes.

With this methodology, two natural questions will arise for this application.

*1. How do we know what objects are interesting to users?* The answer to this question is very subjective and user dependent. Usually people are more interested in faces and texts [33], which requires face and text detection. We could add this function rather easily to the

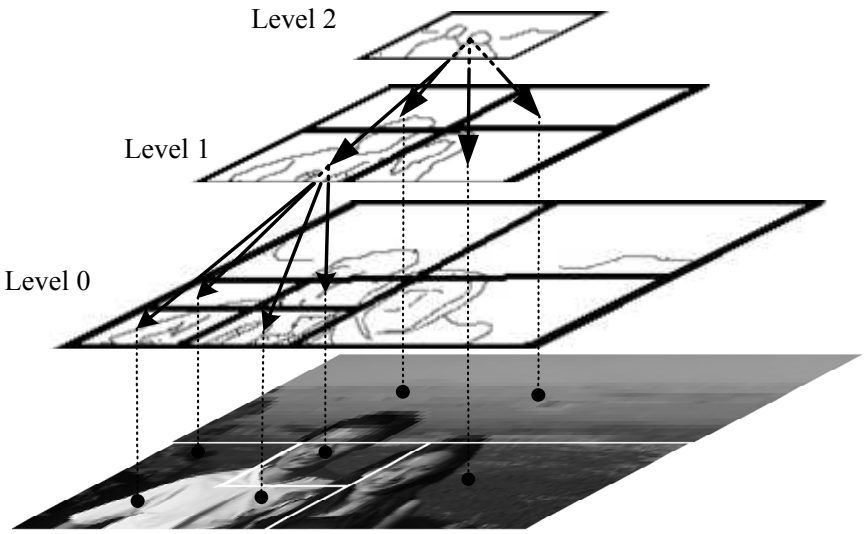



Fig. 21. A Gaussian pyramid is partitioned into a quad-tree. We only show the nodes which are visited by our algorithm. During the "tour", a quad-tree node is visited if and only if its sketch sub-graph expands from the level above, which indicates additional semantic/structural information appears.

system as there are off-the-shelf code for face and text detection working reasonably well. But it is beyond the scope of this paper, which is focused on low-middle level representation of generic images.

*2. How do we measure the information gain when we zoom in an area?* There are two existing criteria: one is the focus of attention models [15], which essentially favors areas of high intensity contrast. A problem with this criterion is that we may be directed to some boring areas, for example smooth (cartoon like) regions where few new structures will be revealed when the area is zoomed-in. The other is to sum up the Fourier power of the Laplacian image at certain scale over a region. This could tell us how much signal power we gain when the area is zoomed in. But the frequency power does not necessarily mean structural information. For example, we may zoom into the forrest (texture) in the background (see Figure 22).

We argue that a sketch pyramid together with perceptual transitions provide a natural and generic measure for the information gains when we zoom in an area or visit a node in

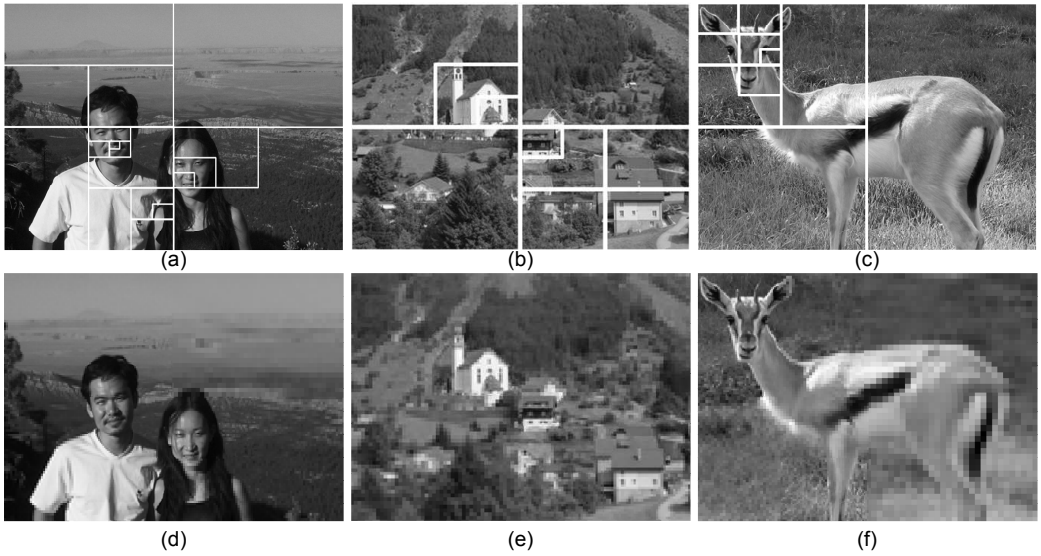



Fig. 22. Images (a-c) show three examples of the tours in the quad-trees. The partitions correspond to regions in the sketch pyramid that experience sketch graph expansions. If the graph expands in a given partition, then we need to increase the resolution of the corresponding image region to capture the added structural information. Images (d-f) represent the replacement of each sketch partition with an image region from the corresponding level in the Gaussian pyramid. Note that the areas of higher structural content are in higher resolution (e.g. face, house), and areas of little structure are in lower resolution (e.g. landscape, grass).

the quadtree.

In computing a perceptual sketch pyramid, we studied the inverse process when we compute operators from high-resolution to low-resolution. A node $v$ at level $k$ corresponds to a sub-graph $\mathbf{S}_k(v)$ of sketches, and its children at the higher resolution level correspond to $\mathbf{S}_{k-1}(v')$. The information gain for this split is measured by

$$\delta(v) = -\log_2 \frac{p(\mathbf{S}_{k-1}(v'))}{p(\mathbf{S}_k(v))}. \tag{7}$$

It is the number of new bits needed to describe the extra information when graph expands. As each node in the quad-tree has an information measure, we can expand a node in a sequential order until a threshold $\tau$ (or a maximum number of bits $M$) is reached. Figure 22 shows results of the quad-tree decomposition and multi-resolution image reconstruction. The

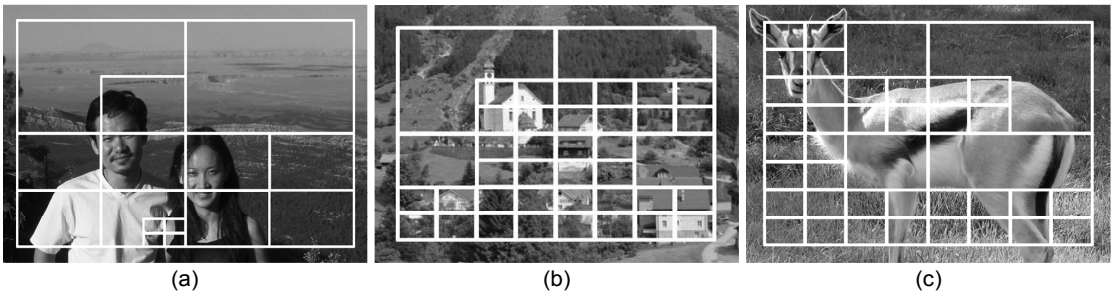
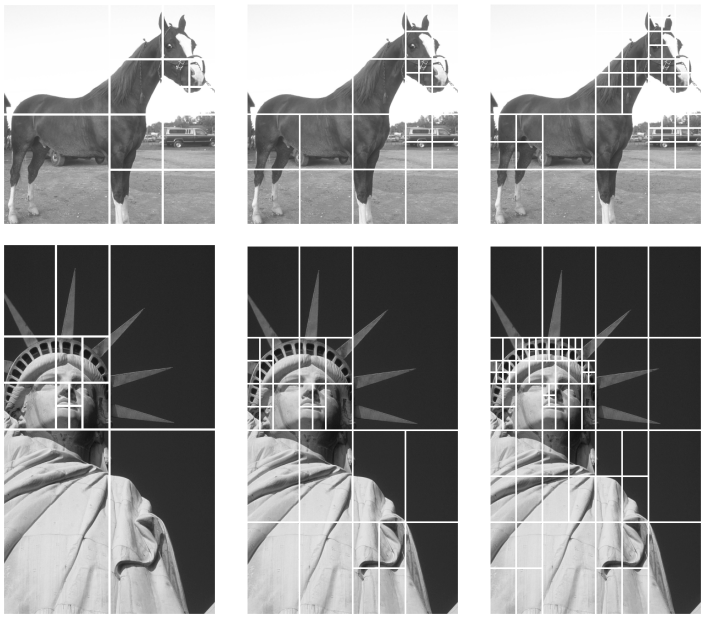



Fig. 23. For comparison with the corresponding partitions in the sketch pyramid (Fig.22), a Laplacian decomposition of the test images are shown. The outer frame is set smaller than the image size to avoid Gaussian filtering boundary effects. The algorithm greedily splits the leaf nodes bearing the most power (sum of squared pixel values in the node of the Laplacian pyramid $I_k^+$). As clearly evident, the Laplacian decomposition does not exploit the perceptually important image regions in its greedy search (e.g. facial features) instead focusing more on the high frequency areas.

reconstructed images show that there is little perceptual loss of information when each region is viewed at its determined scale.

The information gain measure in Eq.7 is more meaningful than calculating the power of bandpass Laplacian images. For example, as shown in Figure 4.(b), a long sharp edge in an image will spread across all levels of the Laplacian pyramid, and thus demands continuous refining in the display if we use the absolute value of the Laplacian image patches. As shown in Figure 5, in contrast, in a sketch pyramid, it is a single edge and will stop at certain high level. As a result, the quad-tree partition makes more sense to human beings in the perceptual sketch pyramid than in the Laplacian of Gaussian pyramid, as shown in Figure 22 and Figure 23. Thus, we argue that perceptual sketch pyramid is more meaningful in representing human perception than Laplacian of Gaussian pyramid adopted by conventional scale space studies.

To evaluate the quad-tree decomposition for images based on the inferred perceptual sketch pyramids, we design the following experiments to quantitatively verify the perceptual sketch pyramid model and its computation.



(a) (b) (c)

Fig. 24. Comparison of the quad-trees partitioned by the human subjects and the computer algorithm – Part II. The computer performance is within the variation of human performance. (a)The computer algorithm partitioned quad-tree. (b) & (c) Two examples of human partitioned quad-tree.

We selected 9 images from the Corel image database and 7 human subjects with and without computer vision background. To get a fair evaluation of the inference results, we deliberately chose 7 different people from the 7 graduate students who had done the perceptual transition labeling in the learning stage. Each human subject was provided with a computer graphical user interface, which allowed them to partition the given high-resolution images into quad-trees as shown in Figure 25. The depth ranges of the quad-trees were specified by the authors in advance. For example, for the first three images in Figure 25, the quad-tree depth range was set to 1 to 4 levels, 1 to 4 levels and 2 to 5 levels, respectively. Then, the human subjects partitioned the given images into quad-trees based on the *information* distributed on the images according to their own perception and within the specified depth ranges.

After collecting the experiment results from human subjects, we compared the quad-tree

partition performance by human subjects and that by our computer algorithm based on inferred perceptual sketch pyramids. The quantitative measure was computed as follows. Each testing image was split into quad-tree by the 7 human subjects. Consequently, each pixel of the image was assigned 7 numbers, which were the corresponding depth of the quad-tree partitioned by the 7 human subjects at the pixel. In Table I, the middle column is the average standard deviation (STD) of quad-tree depth per pixel of each image partitioned by the human subjects. It tells the human performance variation. For each image, we take the average depth of the human partition as the "truth" for each pixel. The right column shows the computer algorithm's partition error from the "truth". This table shows that in 7 out of 9 images, the computer performance is within the variation of human performance.

Figure 25 and Figure 24 compare the partition results between computer algorithm and human subjects. In these figures, the first column shows the quad-tree partitions of each image by our computer algorithm. The other two columns are two sample quad-tree partitions by the human subjects. In Figure 25, our computer algorithm's performance is within the variation of human performance. Figure 24 shows the two exceptional cases. However, from the figure, we can see that the partitions by our computer algorithm are still very reasonable.

## VIII. Summary and discussion

In this paper, we propose a perceptual scale-space representation to account for perceptual jumps amid continuous intensity changes. It is an augmentation to the classical scale-space theory. We model perceptual transitions across scales by a set of context sensitive grammar rules, which are learned through a supervised learning procedure. We explore the mechanism for perceptual transitions. Based on inferred sketch pyramid, we define "information gain" of an image across scales as the number of extra bits needed to describe graph expansion. We show an application of such an information measure for adaptive image display.

Our discussion is mostly focused on the mild perceptual transitions. We have not discussed

| Image ID | Human Partition STD | Computer Partition Error |
|---|---|---|
| 54076 | 0.231265 | 0.202402 |
| 77016 | 0.140912 | 0.092913 |
| 180088 | 0.246501 | 0.154938 |
| 234007 | 0.239068 | 0.208406 |
| 404028 | 0.357495 | 0.325235 |
| 412037 | 0.178133 | 0.162965 |
| street | 0.279815 | 0.263395 |
| 197046* | 0.168721 | 0.244858 |
| 244000* | 0.245081 | 0.369920 |

TABLE I

COMPARISON TABLE OF QUAD-TREE PARTITION PERFORMANCE ON 9 IMAGES BETWEEN HUMAN SUBJECTS AND THE COMPUTER ALGORITHM BASED ON INFERRED PERCEPTUAL IMAGE PYRAMIDS. THIS TABLE SHOWS THAT FOR 7 OUT OF 9 IMAGES, THE COMPUTER PERFORMANCE IS WITHIN THE VARIATION OF HUMAN PERFORMANCE.

explicitly the mechanism for the catastrophic texture-texton transitions, which is referred to in a companion paper[32]. In future work, it shall also be interesting to explore the link between the perceptual scale-space to the multi-scale feature detection and object recognition[14], [10], and applications such as super-resolution, and tracking objects over a large range of distances.

## ACKNOWLEDGEMENTS

This work was supported in part by NSF grants IIS-0222967 and IIS-0413214, and an ONR grant N00014-05-01-0543. Bahrami is also partially supported by an NSF graduate Research Fellowship. The authors also thank Ziqiang Liu for sharing part of his programming code,

and Dr. Xing Xie and Xin Fan at Microsoft Research Asia for discussions on the adaptive image display work.

## References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", *Proc. Int'l Conf. on Very Large Data Bases*, 1994.

[2] T. Chan and J. Shen, "Local Inpainting Model and TV-inpainting," *SIAM J. of Appli. Math*, 62:3, 1019-43, 2001.

[3] T.F. Cootes and G.J. Edwards and C.J. Taylor, "Active Appearance Models," *Proc. European Conf. on Computer Vision*, 1998.

[4] A. Fridman, *Mixed Markov Models*, Doctoral dissertation, Division of Applied Math, Brown University, 2000.

[5] P.J. Green, "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82:711-732, 1995.

[6] C. Guo and S.C. Zhu and Y. Wu, "A Mathematical Theory of Primal Sketch and Sketchability," *Proc. Int'l Conf. on Computer Vision*, 2003.

[7] C. Guo and Y. Wu and S.C. Zhu, "Information Scaling Laws in Natural Scenes," *Prod. 2nd Workshop on Generative Model Based Vision*, 2004.

[8] C.E. Guo, S.C. Zhu and Y.N. Wu, "Primal Sketch: Integrating Texture and Structure," *Computer Vision and Image Understanding (Accepted for the Special Issue on Generative Model Based Vision)*, 2006.

[9] B. Julesz, "Textons, the Elements of Eexture Perception, and Their Interactions," *Nature*, 290:91-97, 1981.

[10] T. Kadir and M. Brady, "Saliency, Scale and Image Description," *Int'l J. Computer Vision*, 2001.

[11] P. Klein and T. Sebastian and B. Kimia, "Shape Matching Using Edit-distance: an Implementation," *SODA*, 781-790, 2001.

[12] J.J. Koenderink, "The Structure of Images," *Biological Cybernetics*, 1984.

[13] T. Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, 1994.

[14] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision*, 2004.

[15] Y.F. Ma , L. Lu , H.J. Zhang and M.J. Li "An Attention Model for Video Summarization", *ACM Multimedia*, 12, 2002.

[16] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[17] S. Mallat and Z. Zhang, "Matching Pursuit in a Time-Frequency Dictionary," *IEEE Trans. on Sigmal Processing*, 41:3397-415, 1993.

[18] D. Marr, *Vision*, Freeman Publisher, 1983.

[19] N. Metropolis, M. Rosenbluth, A. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chemical Physics, 21, 1087-92, 1953.*

[20] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," *Int'l J. Computer Vision*, 2004.

[21] D. B. Mumford and B. Gidas, "Stochastic Models for Generic Images," *Quaterly of Applied Mathematics*, 59(1):85-111, 2001.

[22] D. L. Ruderman, "The Statistics of Natural Images," *Network* 5:517-548, 1994.

[23] B.M. ter Haar Romeny, *Front-End Vision and Multiscale Image Analysis: Introduction to Scale-Space Theory*, Kluwer Academic Publishers, 1997.

[24] E.P. Simoncelli and W.T. Freeman and E.H. Adelson and D.J. Heeger, "Shiftable Multi-scale Transforms," *IEEE Trans. Info. Theory*, 38(2):587-607, 1992.

[25] J. Sporring, M. Nielsen, L. Florack, and P. Johansen, *Gaussian Scale-Space*, Kluwer Academic Publishers, 1996.

[26] A. Srivastava, A.B. Lee, E.P. Simoncelli, and S.C. Zhu, "On Advances in Statistical Modeling of Natural Images," *J. of Math Imaging and Vision*, 18(1):17-33, 2003.

[27] J. Sun, N.N. Zheng, H. Tao, and Y. Y. Shum, "Generic Image Hallucination with Primal Sketch Prior," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[28] Y.A. Wang and E.H. Adelson, "Representing Moving Images with Layers," *IEEE Trans. on Image Processing*, 1994.

[29] Y. Wang and S.C. Zhu, "Modeling Complex Motion by Tracking and Editing Hidden Markov Graphs," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[30] Y. Wang and S. Bahrami and S.C. Zhu, "Perceptual Scale Space and Its Applications," *Proc. Int'l Conf. on Computer Vision*, 2005.

[31] A.P. Witkin, "Scale Space Filtering," *Int'l Joint Conf. on AI* Palo Alto, Kaufman, 1983.

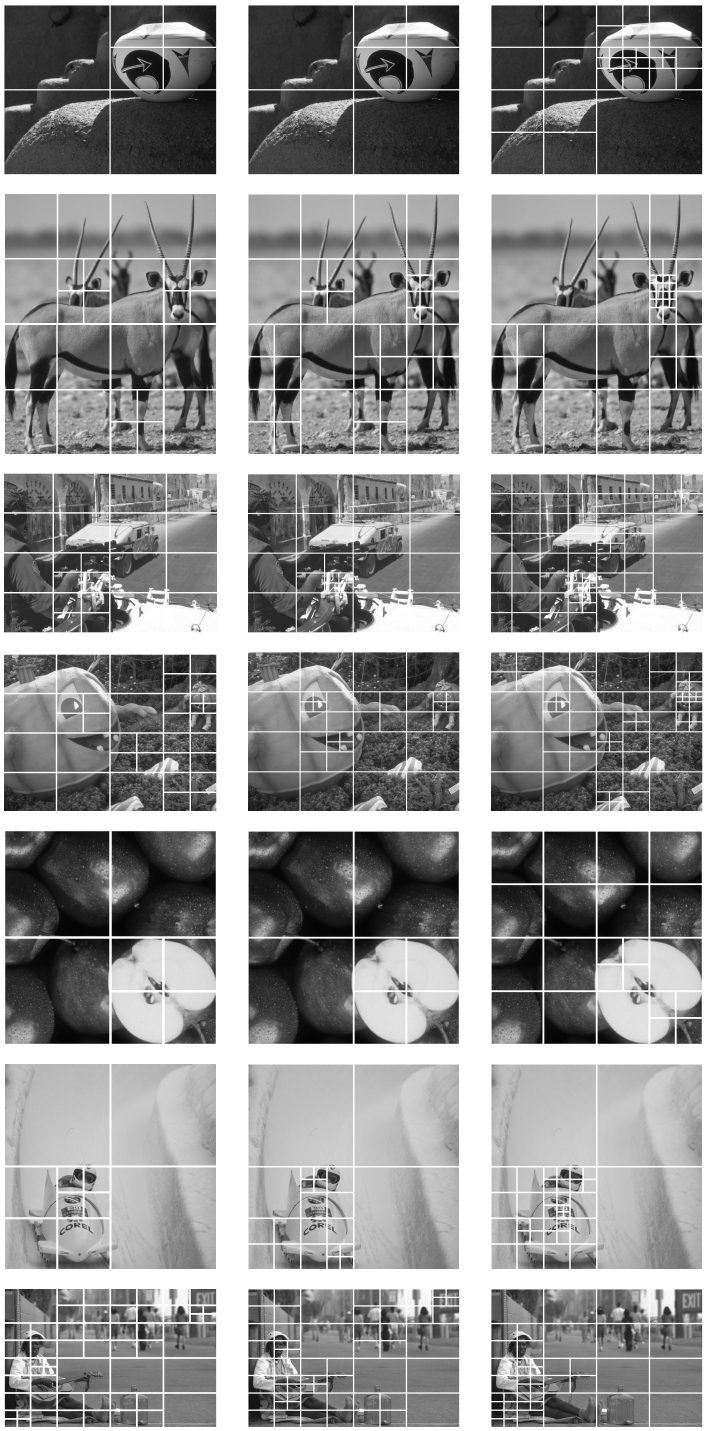
[32] Y.N. Wu, S.C. Zhu, and C.E. Guo, "From Information Scaling Laws of Natural Images to Regimes of Statistical Models," Technical report 408, Department of Statistics, UCLA, 2004.

[33] X. Xie and H. Liu, W.Y. Ma, and H. Zhang, "Browsing Large Pictures Under Limited Display Sizes," *IEEE Trans. on Multimedia*, 2003.

[34] Z.J. Xu, H. Chen and S.C. Zhu, "A High Resolution Grammatical Model for Face Representation and Sketching," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, June 2005.

[35] S.C. Zhu and A.L. Yuille, "FORMS: A Flexible Object Recognition and Modeling System," *Int'l J. Computer Vision*, 20(3):187-212, 1996.

[36] S.C. Zhu and D. Mumford, "Prior Learning and Gibbs Reaction-Diffusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(11):1236-1250, 1997.



(a) (b) (c)

Fig. 25. Comparison of the quad-trees partitions by the human subjects and the computer algorithm. The computer performance is within the variation of human performance. (a)The computer algorithm partitioned quad-tree. (b) & (c) Two examples of human partitioned quad-tree.